(51) International Patent Classification ⁶ :	A1	(11) International Publication Number:	WO 99/37088
H04N 5/262		(43) International Publication Date:	22 July 1999 (22.07.99)

(54) Title: METHOD AND SYSTEM FOR COMBINING VIDEO SEQUENCES WITH SPACIO–TEMPORAL ALIGNMENT

(57) Abstract

Given two video sequences, a composite video sequence can be generated which includes visual elements from each of the given sequences, suitably synchronized and represented in a chosen focal plane. For example, given two video sequences with each showing a different contestant individually racing the same down–hill course, the composite sequence can include elements from each of the given sequences to show the contestants as if racing simultaneously. A composite video sequence can be made also by similarly combining a video sequence with an audio sequence.

METHOD AND SYSTEM FOR COMBINING VIDEO SEQUENCES
WITH SPATIO-TEMPORAL ALIGNMENT

Technical Field

The present invention relates to visual displays and, more specifically,

5   to time-dependent visual displays.

Background of the Invention

In video displays, e.g. in sports-related television programs, special
visual effects can be used to enhance a viewer's appreciation of the action. For
example, in the case of a team sport such as football, instant replay affords the viewer

10   a second chance at "catching" critical moments of the game. Such moments can be
replayed in slow motion, and superposed features such as hand-drawn circles, arrows
and letters can be included for emphasis and annotation. These techniques can be
used also with other types of sports such as racing competitions, for example.

With team sports, techniques of instant replay and the like are most

15   appropriate, as scenes typically are busy and crowded. Similarly, e.g. in the 100-
meter dash competition, the scene includes the contestants side-by-side, and slow-
motion visualization at the finish line brings out the essence of the race. On the other
hand, where starting times are staggered e.g. as necessitated for the sake of
practicality and safety in the case of certain racing events such as downhill racing or

20   ski jumping, the actual scene typically includes a single contestant.

Summary of the Invention

For enhanced visualization, by the sports fan as well as by the
contestant and his coach, displays are desired in which the element of competition
between contestants is manifested. This applies especially where contestants perform

25   sole as in downhill skiing, for example, and can be applied also to group races in
which qualification schemes are used to decide who will advance from quarter-final to
half-final to final.

2

We have recognized that, given two or more video sequences, a composite video sequence can be generated which includes visual elements from each of the given sequences, suitably synchronized and represented in a chosen focal plane. For example, given two video sequences with each showing a different contestant

5   individually racing the same down-hill course, the composite sequence can include elements from each of the given sequences to show the contestants as if racing simultaneously.

A composite video sequence can be made also by similarly combining one or more video sequences with one or more different sequences such as audio

10   sequences, for example.

Brief Description of the Drawing

Fig. 1 is a block diagram of a preferred embodiment of the invention.

Figs. 2A and 2B are schematics of different downhill skiers passing

15   before a video camera.

Figs. 3A and 3B are schematics of images recorded by the video camera, corresponding to Figs. 2A and 2B.

Fig. 4 is a schematic of Figs. 2A and 2B combined.

Fig. 5 is a schematic of the desired video image, with the scenes of Fig.

20   3A and 3B projected in a chosen focal plane.

Fig. 6 is a frame from a composite video sequence which was made with a prototype implementation of the invention.

Detailed Description

Conceptually, the invention can be appreciated in analogy with 2-

25   dimensional (2D) "morphing", i.e. the smooth transformation, deformation or mapping of one image, I1, into another, I2, in computerized graphics. Such morphing leads to a video sequence which shows the transformation of I1 into I2, e.g., of an image of an apple into an image of an orange, or of one human face into another. The video sequence is 3-dimensional, having two spatial and a temporal dimension. Parts

30   of the sequence may be of special interest, such as intermediate images, e.g. the

average of two faces, or composites, e.g. a face with the eyes from I1 and the smile
from I2. Thus, morphing between images can be appreciated as a form of merging of
features from the images.

The invention is concerned with a more complicated task, namely the
5   merging of two video sequences. The morphing or mapping from one sequence to
another leads to 4-dimensional data which cannot be displayed easily. However, any
intermediate combination, or any composite sequence leads to a new video sequence.

Of particular interest is the generation of a new video sequence
combining elements from two or more given sequences, with suitable spatio-temporal
10   alignment or synchronization, and projection into a chosen focal plane. For example,
in the case of a sports racing competition such as downhill skiing, video sequences
obtained from two contestants having traversed a course separately can be time-
synchronized by selecting the frames corresponding to the start of the race.
Alternatively, the sequences may be synchronized for coincident passage of the
15   contestants at a critical point such as a slalom gate, for example.

The chosen focal plane may be the same as the focal plane of the one
or the other of the given sequences, or it may be suitably constructed yet different
from both.

Of interest also is synchronization based on a distinctive event, e.g., in
20   track and field, a high-jump contestant lifting off from the ground or touching down
again. In this respect it is of further interest to synchronize two sequences so that both
lift-off and touch-down coincide, requiring time scaling. The resulting composite
sequence affords a comparison of trajectories.

With the video sequences synchronized, they can be further aligned
25   spatially, e.g. to generate a composite sequence giving the impression of the
contestants traversing the course simultaneously. In a simple approach, spatial
alignment can be performed on a frame-by-frame basis. Alternatively, by taking a
plurality of frames from a camera into consideration, the view in an output image can
be extended to include background elements from several sequential images.

30   Forming a composite image involves representing component scenes in
a chosen focal plane, typically requiring a considerable amount of computerized

processing, e.g. as illustrated by Fig. 1 for the special case of two video input sequences.

Fig. 1 shows two image sequences IS1 and IS2 being fed to a module 11 for synchronization into synchronized sequences IS1' and IS2'. For
5 example, the sequences IS1 and IS2 may have been obtained for two contestants in a down-hill racing competition, and they may be synchronized by the module 11 so that the first frame of each sequence corresponds to its contestant leaving the starting gate.

The synchronized sequences are fed to a module 12 for background-foreground extraction, as well as to a module 13 for camera coordinate transformation
10 estimation. For each of the image sequences, the module 12 yields a weight-mask sequence (WMS), with each weight mask being an array having an entry for each pixel position and differentiating between the scene of interest and the background/foreground. The generation of the weight mask sequence involves computerized searching of images for elements which, from frame to frame, move
15 relative to the background. The module 13 yields sequence parameters SP1 and SP2 including camera angles of azimuth and elevation, and camera focal length and aperture among others. These parameters can be determined from each video sequence by computerized processing including interpolation and matching of images. Alternatively, a suitably equipped camera can furnish the sequence parameters
20 directly, thus obviating the need for their estimation by computerized processing.

The weight-mask sequences WMS1 and WMS2 are fed to a module 13 for "alpha-layer" sequence computation. The alpha layer is an array which specifies how much weight each pixel in each of the images should receive in the composite image.

25 The sequence parameters SP1 and SP2 as well as the alpha layer are fed to a module 15 for projecting the aligned image sequences in a chosen focal plane, resulting in the desired composite image sequence. This is exemplified further by Figs. 2A, 2B, 3A, 3B, 4 and 5

Fig. 2A shows a skier A about to pass a position marker 21, with the
30 scene being recorded from a camera position 22 with a viewing angle $\phi(A)$. The position reached by A may be after an elapse of t(A) seconds from A's leaving the

starting gate of a race event.

Fig. 2B shows another skier, B, in a similar position relative to the marker 21, and with the scene being recorded from a different camera position 23 and with a different, more narrow viewing angle φ(B). For comparison with skier A, the
5   position of skier B corresponds to an elapse of t(A) seconds from B leaving the starting gate. As illustrated, within t(A) seconds skier B has traveled farther along the race course as compared with skier A.

Figs. 3A and 3B show the resulting respective images.

Fig. 4 shows a combination with Figs. 2A and 2B superposed at a
10   common camera location.

Fig. 5 shows the resulting desired image projected in a chosen focal plane, affording immediate visualization of skiers A and B as having raced jointly for t(A) seconds from a common start.

Fig. 6 shows a frame from a composite image sequence generated by a
15   prototype implementation of the technique, with the frame corresponding to a point of intermediate timing. The value of 57.84 is the time, in seconds, that it took the slower skier to reach the point of intermediate timing, and the value of +0.04 (seconds) indicates by how much he is trailing the faster skier.

The prototype implementation of the technique was written in the "C"
20   programming language, for execution on a SUN Workstation or a PC, for example. Dedicated firmware or hardware can be used for enhanced processing efficiency, and especially for signal processing involving matching and interpolation.

Individual aspects and variations of the technique are described below in further detail.

25   A.   Background/Foreground Extraction

In each sequence, background and foreground can be extracted using a suitable motion estimation method. This method should be "robust", for background/foreground extraction where image sequences are acquired by a moving camera and where the acquired scene contains moving agents or objects. Required
30   also is temporal consistency, for the extraction of background/foreground to be stable

6

over time. Where both the camera and the agents are moving predictably, e.g. at

constant speed or acceleration, temporal filtering can be used for enhanced temporal

consistency.

Based on determinations of the speed with which the background

5   moves due to camera motion, and the speed of the skier with respect to the camera,

background/foreground extraction generates a weight layer which differentiates

between those pixels which follow the camera and those which do not. The weight

layer will then be used to generate an alpha layer for the final composite sequence.

B.   Spatio-temporal Alignment of Sequences

10   Temporal alignment involves the selection of corresponding frames in

the sequences, according to a chosen criterion. Typically, in sports racing

competitions, this is the time code of each sequence delivered by the timing system,

e.g. to select the frames corresponding to the start of the race. Other possible time

criteria are the time corresponding to a designated spatial location such as a gate or

15   jump entry, for example.

Spatial alignment is effected by choosing a reference coordinate

system for each frame and by estimating the camera coordinate transformation

between the reference system and the corresponding frame of each sequence. Such

estimation may be unnecessary when camera data such as camera position, viewing

20   direction and focal length are recorded along with the video sequence. Typically, the

reference coordinate system is chosen as one of the given sequences— the one to be

used for the composite sequence. As described below, spatial alignment may be on a

single-frame or multiple-frame basis.

B.1   Spatial Alignment on a Single-frame Basis

25   At each step of this technique, alignment uses one frame from each of

the sequences. As each of the sequences includes moving agents/objects, the method

for estimating the camera coordinate transformation needs to be robust. To this end,

the masks generated in background/foreground extraction can be used. Also, as

motivated for background/foreground extraction, temporal filtering can be used for

7

enhancing the temporal consistency of the estimation process.

B.2   Spatial Alignment on a Multiple-frame Basis

        In this technique, spatial alignment is applied to reconstructed images of the scene visualized in each sequence. Each video sequence is first analyzed over multiple frames for reconstruction of the scene, using a technique similar to the one for background/foreground extraction, for example. Once each scene has been separately reconstructed, e.g. to take in as much background as possible, the scenes can be spatially aligned as described above.

        This technique allows free choice of the field of view of every frame in the scene, in contrast to the single-frame technique where the field of view has to be chosen as the one of the reference frame. Thus, in the multiple-frame technique, in case that all contestants are not visible in all the frames, the field and/or angle of view of the composite image can be chosen such that all competitors are visible.

C.   Superimposing of Video Sequences

        After extraction of the background/foreground in each sequence and estimation of the camera coordinate transformation between each sequence and a reference system, the sequences can be projected into a chosen focal plane for simultaneous visualization on a single display. Alpha layers for each frame of each sequence are generated from the multiple background/foreground weight masks. Thus, the composite sequence is formed by transforming each sequence into the chosen focal plane and superimposing the different transformed images with the corresponding alpha weight.

D.   Applications

        Further to skiing competitions as exemplified, the techniques of the invention can be applied to other speed/distance sports such as car racing competitions and track and field, for example.

        Further to visualizing, one application of a composite video sequence made in accordance with the invention is apparent from Fig. 6, namely for

8

determining differential time between two runners at any desired location of a race. This involves simple counting of the number of frames in the sequence between the two runners passing the location, and multiplying by the time interval between frames.

5   A composite sequence can be broadcast over existing facilities such as network, cable and satellite TV, and as video on the Internet, for example. Such sequences can be offered as on-demand services, e.g. on a channel separate from a strictly real-time main channel. Or, instead of by broadcasting over a separate channel, a composite video sequence can be included as a portion of a regular

10   channel, displayed as a corner portion, for example.

           In addition to their use in broadcasting, generated composite video sequences can be used in sports training and coaching. And, aside from sports applications, there are potential industrial applications such as car crash analysis, for example.

15   It is understood that composite sequences may be higher-dimensional, such as composite stereo video sequences.

           In yet another application, one of the given sequences is an audio sequence to be synchronized with a video sequence. Specifically, given a video sequence of an actor or singer, A, speaking a sentence or singing a song, and an audio

20   sequence of another actor, B, doing the same, the technique can be used to generate a voice-over or "lip-synch" sequence of actor A speaking or singing with the voice of B. In this case, which requires more than mere scaling of time, dynamic programming techniques can be used for synchronization.

           The spatio-temporal realignment method can be applied in the

25   biomedical field as well. For example, after orthopedic surgery, it is important to monitor the progress of a patient's recovery. This can be done by comparing specified movements of the patient over a period of time. In accordance with an aspect of the invention, such a comparison can be made very accurately, by synchronizing start and end of the movement, and aligning the limbs to be monitored

30   in two or more video sequences.

Another application is in car crash analysis. The technique can be used for precisely comparing the deformation of different cars crashed in similar situations, to ascertain the extent of the difference. Further in car crash analysis, it is important to compare effects on crash dummies. Again, in two crashes with the same type of car, one can precisely compare how the dummies are affected depending on configuration, e.g. of safety belts.

CLAIMS

1. A method for generating a composite video sequence from a plurality of given video sequences, comprising:

      (a) synchronizing the given sequences; and

      (b) forming the composite sequence from the synchronized sequences as projected into a chosen focal plane.

2. The method according to claim 1, wherein synchronizing is with respect to a timed event in the given sequences.

3. The method according to claim 1, wherein synchronizing is with respect to a common spatial event in the given sequences.

4. The method according to claim 1, wherein synchronizing is with respect to two events in each of the given sequences, with time scaling for equalizing time between the events.

5. The method according to claim 1, wherein the given video sequences have camera parameters including camera location and focal length, wherein the chosen focal plane corresponds to focal plane of the one of the given sequences, and wherein the composite sequence is as viewed from the camera location of the one of the given sequences.

6. The method according to claim 1, wherein forming the composite sequence is on a frame-by-frame basis.

7. The method according to claim 1, wherein forming the composite sequence is based on several frames of at least one of the sequences, for an expanded field of view in the composite sequence as compared with the one of the sequences.

8. The method according to claim 1, wherein the given video sequences are from a sports event.

9. The method according to claim 8, wherein the sports event is a ski race.

10. The method according to claim 8, wherein the sports event is a car race.

5   11. A system for generating a composite video sequence from a plurality of given video sequences, comprising:

        (a) means for synchronizing the given sequences and

        (b) means for forming the composite sequence from the synchronized sequences as projected into a chosen focal plane.

10   12. The system according to claim 11, wherein the means for synchronizing comprises means for aligning the given sequences with respect to a timed event in the given sequences.

13. The system according to claim 11, wherein the means for synchronizing comprises means for aligning the given sequences with respect to a common spatial

15   event in the given sequences.

14. The system according to claim 11, wherein the means for synchronizing comprises means for aligning the given sequences with respect to two events in each of the given sequences, and means for time scaling for equalizing time between the events.

20   15. The system according to claim 11, wherein the given video sequences have camera parameters including camera location and focal length, and wherein the means for forming the composite sequence comprises means for choosing the focal plane corresponding to focal plane of the one of the given sequences, and means for forming the composite sequence as viewed from the camera location of the one of the

given sequences.

16. The system according to claim 11, wherein the means for forming the composite sequence comprises means for processing the given sequences on a frame-by-frame basis.

5   17. The system according to claim 11, wherein the means for forming the composite sequence comprises means for using several frames of at least one of the sequences, for an expanded field of view in the composite sequence as compared with the one of the sequences.

18. The system according to claim 11, wherein the given video sequences are 10   from a sports event.

19. The system according to claim 18, wherein the sports event is a ski race.

20. The system according to claim 18, wherein the sports event is a car race.

21. A system for generating a composite video sequence from a plurality of given video sequences, comprising:

15   (a) means for synchronizing the given sequences and

        (b) a processor which is instructed for forming the composite sequence from the synchronized sequences as projected into a chosen focal plane.

22. A method for generating a composite image, comprising:

        (a) synchronizing a plurality of video sequences and

20   (b) forming the composite image from corresponding frames of the synchronized sequences as projected into a chosen focal plane.

23. A broadcast service comprising:

        (a) synchronizing a plurality of video sequences;

(b) forming a composite sequence from the synchronized sequences as projected into a chosen focal plane; and

(c) broadcasting the composite sequence.

24. A method for generating a composite video sequence from a given video sequence and a given audio sequence, comprising:

(a) synchronizing the given video sequence and the given audio sequence for synchrony between visual features of the video sequence and audio features of the audio sequence; and

(b) forming the composite sequence from the synchronized sequences, having a video portion corresponding to the given video sequence and an audio portion corresponding to the given audio sequence.

25. A system for generating a composite video sequence from a given video sequence and a given audio sequence, comprising:

(a) means for synchronizing the given video sequence and the given audio sequence for synchrony between visual features of the video sequence and audio features of the audio sequence; and

(b) means for forming the composite sequence from the synchronized sequences, having a video portion corresponding to the given video sequence and an audio portion corresponding to the given audio sequence.

26. A system for generating a composite video sequence from a given video sequence and a given audio sequence, comprising:

(a) means for synchronizing the given video sequence and the given audio sequence for synchrony between visual features of the video sequence and audio features of the audio sequence; and

(b) a processor which is instructed for forming the composite sequence from the synchronized sequences, having a video portion corresponding to the given video sequence and an audio portion corresponding to the given audio sequence.

27. A method for determining differential time between two contestants at a specified location in a race, comprising:

forming a composite video sequence from synchronized given video sequences of the contestants, projected into a chosen focal plane; and

5   counting the number of frames between the contestants passing the location.

28. The method according to claim 1, wherein the given video sequences have biomedical significance.

29. The method according to claim 28, wherein biomedical significance

10   comprises significance as to movement of a limb of a patient.

30. The method according to claim 1, wherein the given video sequences comprise car crash test sequences.

31. The method according to claim 30, wherein the car crash test sequences comprise images of cars being tested.

15   32. The method according to claim 30, wherein the car crash test sequences comprise images of crash dummies in cars being tested.
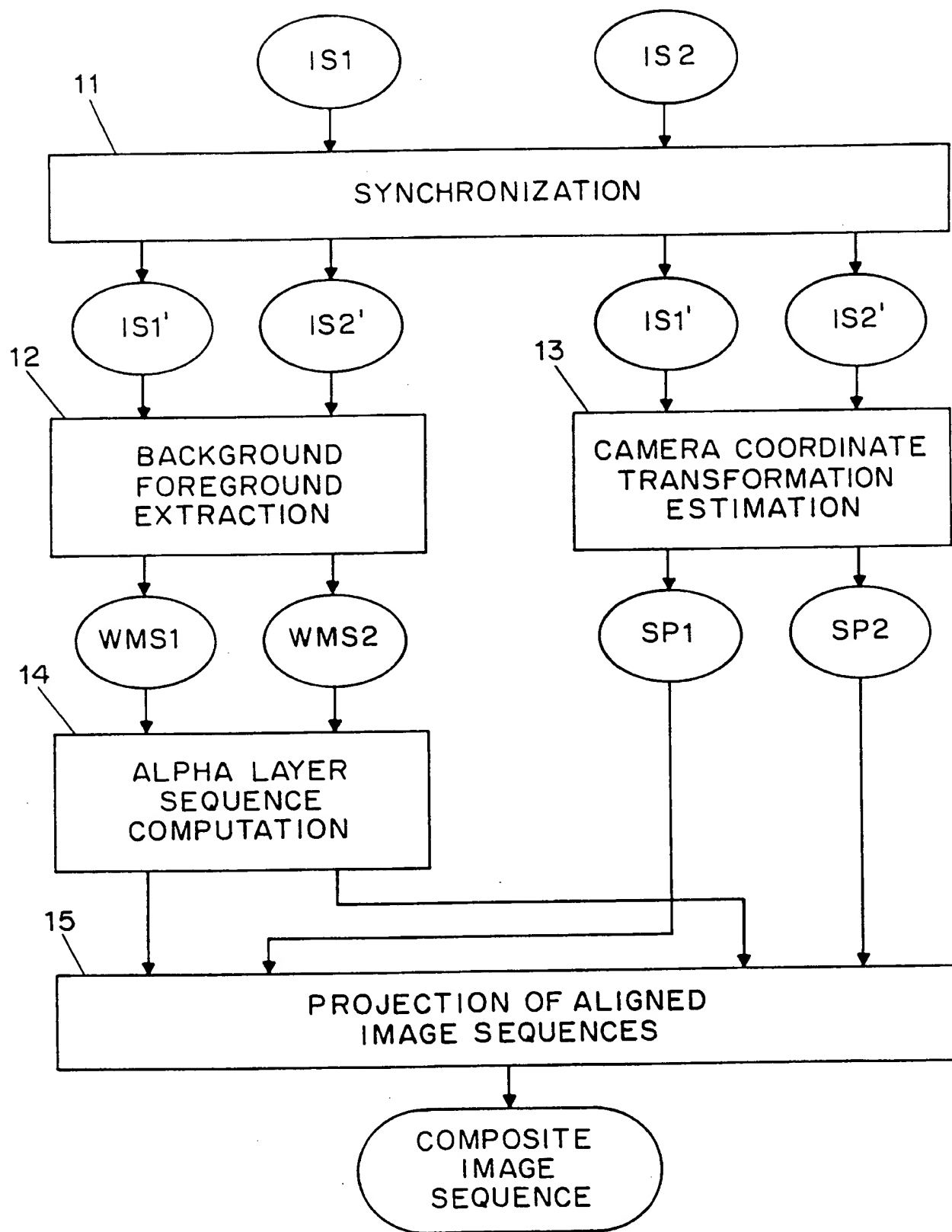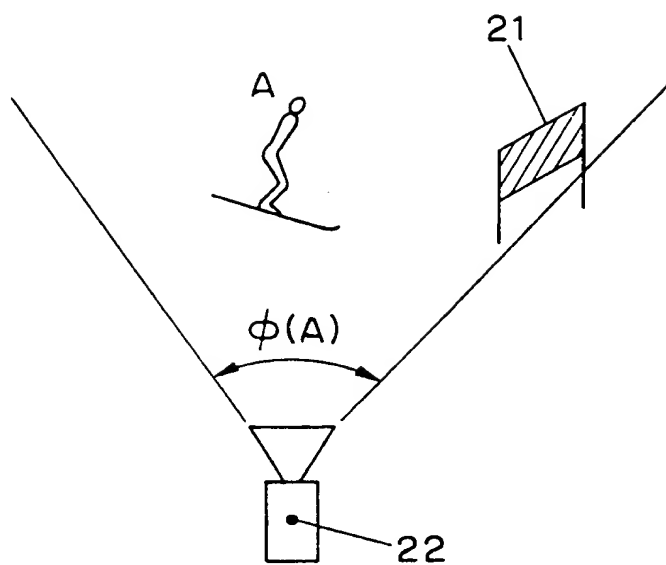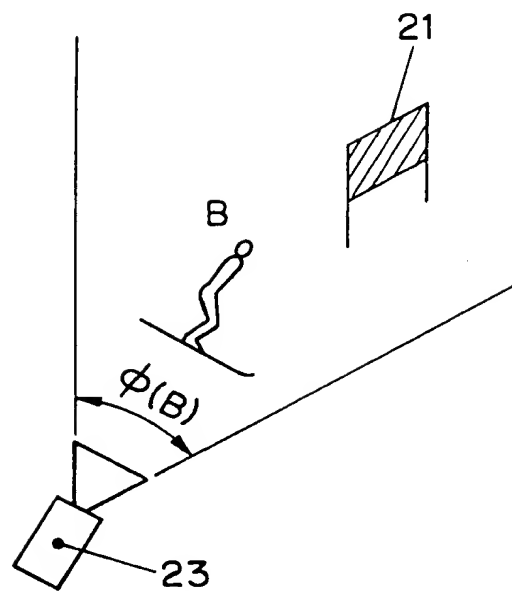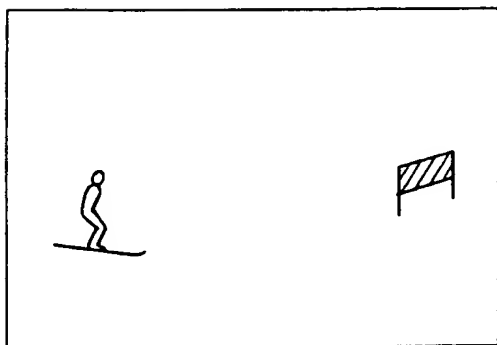
FIG. 1

FIG. 2A

FIG. 2B

FIG. 3A
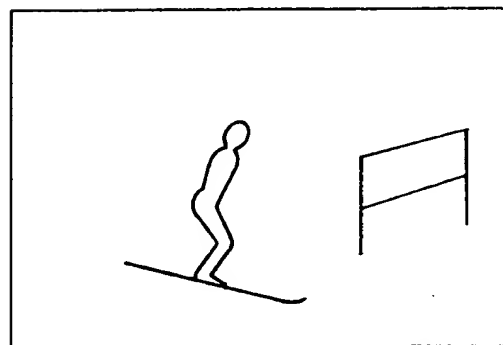
FIG. 3B
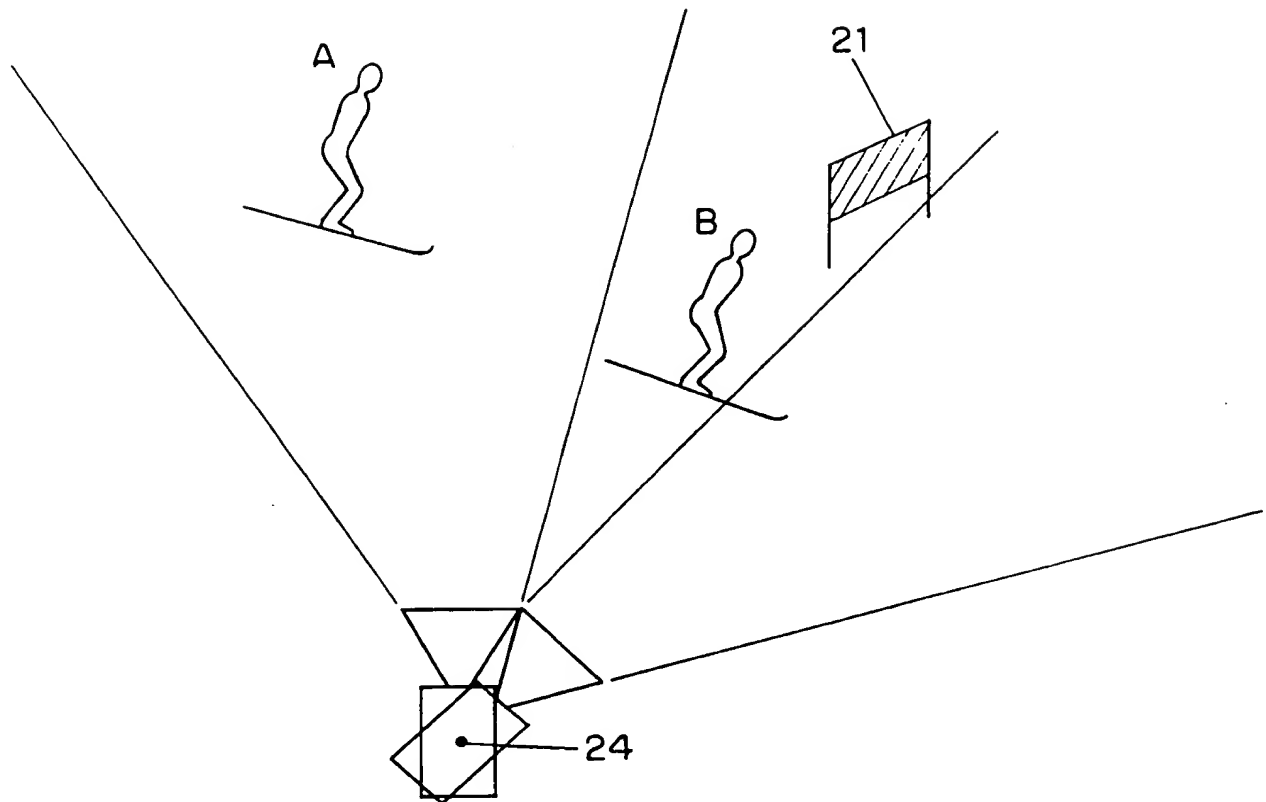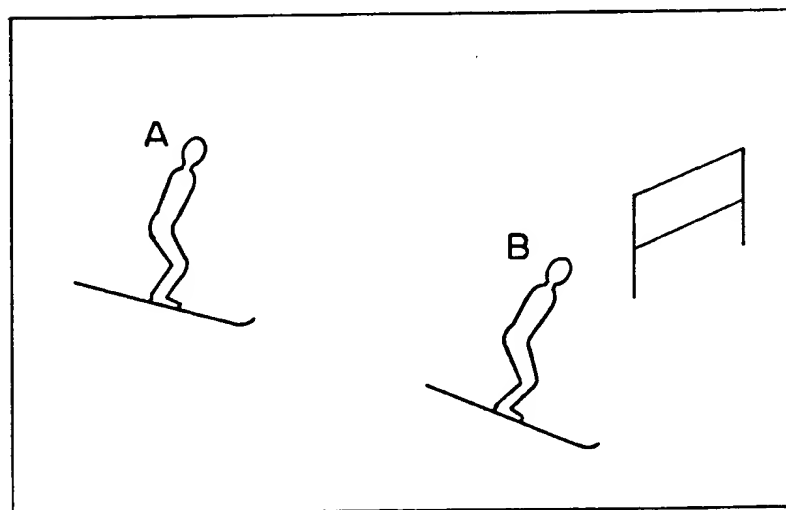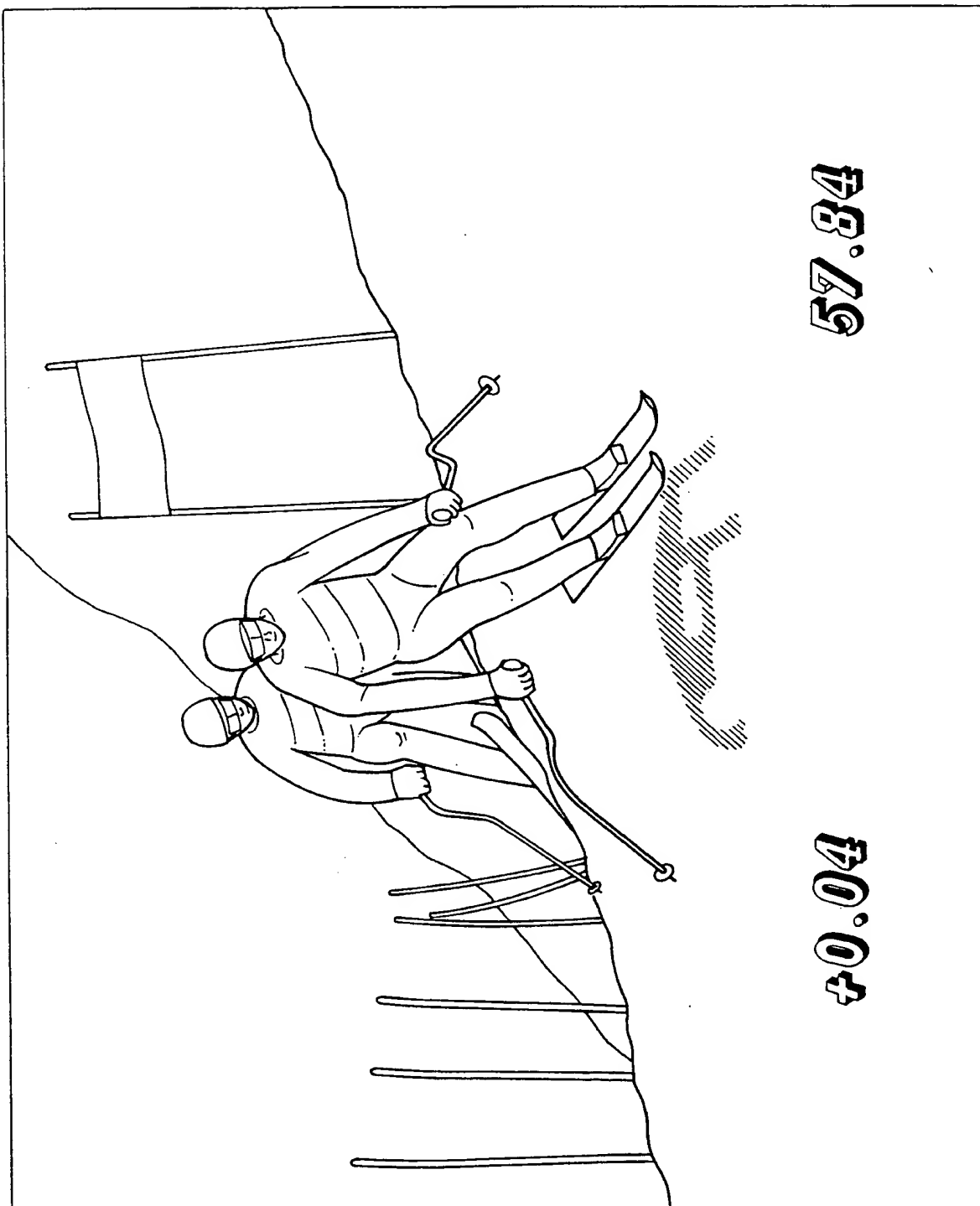
FIG. 4

FIG. 5

FIG. 6

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6   H04N5/262

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6   H04N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EP 0 669 758 A (OH YOH KEISOKU KENKYUSHO KK) 30 August 1995 see column 30, line 9 – line 38 see column 48, line 39 – column 49, line 52 see figures 1,2,12,35 ---	1-27
A	WO 95 05873 A (EASTERBROOK NORMAN J) 2 March 1995 see page 4, line 32 – page 5, line 10 ---	1,8-11, 18-20, 28,29
A	US 4 843 483 A (BOGNER WILLY) 27 June 1989 ---	
A	US 5 367 286 A (VOILLAT JEAN-PIERRE) 22 November 1994 -----	

☐ Further documents are listed in the continuation of box C.   ☒ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not In conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
4 June 1999	10/06/1999

Name and mailing address of the ISA	Authorized officer
European Patent Office. P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040. Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Wentzel, J

Form PCT/ISA/210 (second sheet) (July 1992)

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
EP 0669758	A	30-08-1995	JP	7079382 A	20-03-1995
			JP	7162744 A	23-06-1995
			JP	7184115 A	21-07-1995
			WO	9507590 A	16-03-1995
WO 9505873	A	02-03-1995	AU	682528 B	09-10-1997
			AU	7377794 A	21-03-1995
			CN	1139388 A	01-01-1997
			EP	0746390 A	11-12-1996
			JP	9506978 T	08-07-1997
			US	5823786 A	20-10-1998
US 4843483	A	27-06-1989	DE	3622618 C	21-05-1987
			AT	102418 T	15-03-1994
			CA	1327847 A	15-03-1994
			DE	3789169 D	07-04-1994
			EP	0252215 A	13-01-1988
			ES	2050101 T	16-05-1994
			JP	1920422 C	07-04-1995
			JP	6052934 B	06-07-1994
			JP	63027172 A	04-02-1988
US 5367286	A	22-11-1994	CH	681930 A	15-06-1993
			FR	2676569 A	20-11-1992
			WO	9217862 A	15-10-1992
			EP	0533861 A	31-03-1993